# On Evolutionary Game Theory and Team Reasoning*

Daniel Lempert†

Evolutionary game theory has a lengthy history of modeling human interactions, and has been recently used to analyze the emergence and long-term viability of team reasoning. I review some basic elements of evolutionary analysis, and discuss a few issues attending evolutionary game theory's importation from biology (where it was originally used to study genetic evolution of animal behavior) to the human sciences; in particular, I emphasize important differences between genetic and cultural evolution. After sketching a few fundamental results, I describe recent evolutionary analyses of team reasoning. Finally, I suggest some open lines of theoretical and empirical inquiry.

***evolutionary game theory – team reasoning – cultural evolution***

## *Théorie des jeux évolutionnaire et raisonnement en équipe*

La théorie des jeux évolutionnaires a été utilisée de façon récurrente pour modéliser les interactions humaines, et a été plus récemment appliquée pour analyser l'émergence et la viabilité à long terme du raisonnement en équipe. Je présente les bases de l'analyse évolutionnaire, et discute plusieurs problèmes liés à l'importation de la théorie des jeux évolutionnaires depuis la biologie (où elle a été initialement utilisée pour étudier l'évolution génétique des comportements animaux) vers les sciences humaines. Je souligne en particulier des différences importantes entre l'évolution génétique et l'évolution culturelle. Après avoir présenté quelques résultats fondamentaux de l'analyse évolutionnaire, je discute des analyses évolutionnaires récentes du raisonnement en équipe. Je conclus en suggérant plusieurs pistes de recherche théoriques et empiriques.

**théorie des jeux évolutionnaire – raisonnement en équipe – évolution culturelle**

*JEL codes: C73, D91*

# 1. Introduction

When a player team reasons, she identifies the collective strategy profile that best promotes her group's interest, and then plays her part of this

* I thank the editor of this issue and two anonymous reviewers for helpful suggestions.
† Assistant Professor of Politics, SUNY Potsdam. email: lemperds@potsdam.edu.

collective profile. In contrast, the individual reasoner, the conventional rational actor, acts to maximize only his own individual payoffs. The appeal of team reasoning can be most readily appreciated by considering two simple games that seem to present a puzzle for individual reasoning based "orthodox" decision theory (*e.g.*, Bacharach [2006], 35-68; Gold and Sugden [2007], 281-285). First, consider the Hi Lo (see Table 5, Section 8). The intuitively compelling choice is Hi, and (Hi, Hi) is the Pareto-optimal equilibrium. But individual reasoning does not require this choice. A rational player (one who acts to maximize his expected individual payoffs) playing another rational player, under the assumption that the players' rationality is common knowledge, is only entitled to conclude that *I should choose Hi if my opponent selects Hi, and Lo if he chooses Lo* (Gold and Sugden [2007], 284). Second, consider the Prisoners' Dilemma (PD) (see Table 2). It is well-known that individual reasoning (individual payoff maximization) mandates the choice of D; yet, many people have the intuition that C is the correct choice. Team reasoning justifies these intuitions.[1]

Bacharach [1999, 2006] and Sugden [1993, 2000] propose team reasoning as an alternative account of how people make decisions when interacting with others. (There is also a related, though not overlapping, literature in philosophy, including Gilbert [1989], Hakli, Miller and Tuomela [2010] Hollis [1998], Hurley [1989], and Regan [1980].) When a player team reasons, instead of asking (as in the standard account), "what should *I* (as an individual) do?" she asks, "what should *we* (as a team) do?" (see, *e.g.* Gold and Sugden [2007], 285). She answers the latter question by "work[ing] out the best feasible combination of actions for all members of her team" (Bacharach [2006], 111). It is convenient (though not strictly required by the theory) to make the simplifying assumption that the "best feasible combination of actions" is that which leads to the outcome maximizing the sum of team members' individual payoffs. Finally, she takes the action that the "best feasible combination of actions" requires of her; in other words, she chooses the strategy prescribed for her in the team utility maximizing strategy profile.

Team reasoners interacting with each other unambiguously choose Hi in the Hi Lo, and choose Cooperate (C) in the PD, since the combination of actions that maximizes combined individual payoffs is (Hi, Hi) in Hi Lo, and (C, C) in the PD. Thus, team reasoning solves, in the Hi Lo, an equilibrium selection problem that is theoretically problematic for individual reasoning; in the PD, it leads to an outcome that is Pareto-preferable to the (D, D) equilibrium outcome that results when individual reasoners play.

The literature now includes substantial work on the theory of team reasoning (*e.g.*, Bacharach [1999]; Bacharach [2006]; Smerilli [2012]; Sugden [1993]; Sugden [2000]) and a rapidly burgeoning set of articles presenting empirical evidence in favor of team reasoning. The empirical literature now includes a notable number of experiments with strong internal validity that test team reasoning against not only individual reasoning, but against vari-

1. Parts of this section draw substantially on Amadae and Lempert [2015, Sec. 2].

ous behavioral theories that seek to explain interactive choice and cooperation in social dilemmas. Considering the sum total of evidence, it is fair to say that team reasoning fares at least as well as competing theories. A very brief review of key and recent studies follows.

Colman, Pulford and Rose [2008] employs a set of 3×3 and 5×5 games to consider whether subjects choose consistently with equilibria implied by team reasoning or with individual reasoning-implied Nash equilibria. More often, equilibria implied by team reasoning are selected, although vignettes that are intended to prompt (respectively) individual and team reasoning have notable effects, suggesting that the mode of reasoning employed can be affected by external primes. Bardsley, Mehta, Starmer and Sugden [2010] presents two sets of experiments that test whether subjects playing coordination games use team reasoning or level-k reasoning[2] to make decisions; both explanations receive some support. Butler [2012] gives experimental evidence based on various 2×2 games for the proposition that people vary in their inclination toward team reasoning, and that whether they do so in a given instance is a function, in part, of the *individual* benefits associated with "cooperating" (*i.e.*, team reasoning on the assumption that the other player is also a team reasoner) rather than "defecting" (engaging in individual reasoning). To distinguish between team reasoning, level-k reasoning (see note 2) and strong Stackelberg reasoning,[3] Colman, Pulford and Lawrence [2014] presents experimental subjects with 4×4 coordination games wherein each choice is uniquely associated with one of the four theories. The results are quite mixed, but (a modified version of) level-1 reasoning, team reasoning, and strong Stackelberg reasoning all receive some support. Both Faillo, Smerilli and Sugden [2017] and Bardsley and Ule [2017] assess the relative strength of evidence supporting level-k and team reasoning in coordination games. The former considers a series of 3×3 coordination games, and finds that team reasoning predicts choices quite well, except when the equilibrium implied by team reasoning is payoff dominated by two equilibria that are isomporphic to each other; there is interesting suggestive evidence for a form of boundedly rational ("naive") team reasoning. The latter discusses an experiment wherein subjects play 10×10 risky coordination games of varying complexity; the evidence favors team reasoning – interestingly, players do not appear to engage in level-1 reasoning even when explicitly told that they are faced with a computer that chooses randomly (*i.e.*, a level-0 reasoner). Finally, Pulford, Colman, Lawrence and Krockow [2017] evaluates the explanatory power of various theories of cooperation with an experiment involving several variations on the Centipede

2. Roughly, the theory of level-k reasoning proposes that each individual reasons at a given level about the reasoning of others. A level-0 player chooses randomly; a level-1 player chooses a best-response conditional on the assumption that all players that she interacts with are level-0 players; a level-2 player's best response is conditioned on his assumption that he encounters a mix of level-0 and level-1 players, and so on. For one overview, see Bardsley *et al.* [2010, 43-47].

3. In essence, a strong Stackelberg reasoner chooses his strategy on the assumption that his co-player can perfectly anticipate the strong Stackelberg reasoner's choice, and therefore that the co-player will play a best response to that choice (see *e.g.*, Colman, Pulford and Lawrence [2014], 43-45).

game.[4] Team reasoning outperforms several other theories, including (intra-game) reciprocity, but fuzzy trace theory[5] is also substantially supported.

Despite its sound theoretical basis and the body of laboratory evidence in support, there remains one additional ground for skepticism about team reasoning to address. The objection stems from the observation that team reasoning seems to require behavior that is potentially self-sacrificial. Thus, one may doubt whether behavior implied by team reasoning is viable in the long run. Similarly, one may object to team reasoning as fundamentally *individually* irrational. Evolutionary models of team reasoning have addressed these issues and come to conclusions that are more positive than initial considerations might suggest.

Still, there is room for additional theoretical work on evolutionary models of team reasoning. I argue below that the means by which modes of reasoning are transmitted should be given more careful attention. Specifically, I explain why it is likely that team reasoning is (in the main) culturally, not genetically, transmitted. As such, I suggest, the standard models of evolutionary game theory may need to be modified to give a fully satisfactory evaluation of team reasoning's viability.

In this article, I aim to provide some basic background on evolutionary models of social interaction, with a focus on issues relevant to team reasoning. Early, and still today orthodox, evolutionary analyses of human interaction were of course formulated before team reasoning was formalized (Bacharach [1999]; Sugden [1993]). Still, these models are important for at least two reasons. First, they were taken to shed light on the character and degree of human cooperation in a very general sense; it is straightforward that this has implications for team reasoning's *explanandum*. Second, subsequent evolutionary analyses can be understood as a response to these theoretically compelling models' shortcomings when weighed against real-world empirical observation and experimental data. After giving this background, I describe recent evolutionary analyses of team reasoning, and finally suggest a few open lines of theoretical and empirical inquiry that these models suggest.

A holistic inquiry into questions raised here implicates, at a minimum, literature from theoretical and behavioral economics, biology, genetics, anthropology, and psychology. Though I have attempted incorporate the most centrally relevant information from each of these fields, the review here is necessary selective in both breadth and depth. I keep the discussion as non-technical as feasible, even at the cost of some precision; readers may wish to refer to the literature cited for certain specifics.

4. In its basic form, the Centipede game is a two-player extensive form game wherein the players alternate over a large (fixed and known) number of decision nodes. At each node, a player can choose to continue or to end the game. The payoff from ending the game is (a) higher than continuing given that the other player ends the game in her next move but (b) lower than continuing given that the other player also continues in her next move. The game is famous as a critique of backward induction, which predicts – contra empirical observation – that the game will be ended at the first choice node.

5. Very roughly, fuzzy trace theory posits that decision-making involving numerical calculations is made by the crudest heuristic that allows choices to be ranked. (see *e.g.* Pulford *et al.* [2017], 107).

In Section 2, I describe some basic elements of, and modeling choices associated with, evolutionary models. Section 3 discusses issues in interpreting evolutionary models, in particular the distinction between genetic and cultural evolution. In Section 4, I describe a few fundamental concepts and results in the literature, having to do with evolutionary stability and group selection. Section 5 sketches theoretical work probing the evolutionary dynamics of team reasoning. Finally, Section 6 suggests avenues of future research implicated by the work reviewed here.

# 2. Elements of Evolutionary Models

A wide range of theoretical approaches can be subsumed under the umbrella of *evolutionary models*. The relevant commonality among the approaches is the assumption that, within a given population, the strategies of interactants that perform relatively well in terms of some objective payoff at some time $t$ will be relatively better-represented at the next time period $t+1$. Thus, evolutionary models are suited to assess long-term prevalence of strategies ("types" or "traits," depending on the context), given this key assumption. For example, consider a population from which, in each time period, players are randomly drawn to play the (trivial) game in Table 1. Suppose the two types are "All-A" (*i.e.*, always play $A$) and "All-B" *(i.e.*, always play $B$). It is easy to see that All-A does relatively better than All-B, whether interacting with an All-A or with an All-B type. Thus, whatever the initial proportion of the two types in the population, in each time period the proportion of All-A types increases, so that in the long-run only All-A will be present. Of course, the conclusions of evolutionary models in the literature are rarely so simple, and the implications of various models models can appear (at least initially) to be contradictory. To better understand some of these apparent inconsistencies, it is useful to mention a few important ways in which evolutionary models can differ, and suggest some implications for cooperation or efficient outcomes.

First, the structure of the game that is played can vary. Commonly (if not usually), the PD (Table 2) is analyzed, following Trivers [1971]; see Skyrms [2003, 1-14] for critical discussion. Another important game is Hawk-Dove in Table 3 (Smith and Price [1973]), which, when $v < c$, is identical to the classic game of Chicken. Skyrms [2003] argues that it is the Stag Hunt of Table 4 that captures the essence of human interactions – in particular, formation and maintenance of a "social contract."[6] Among others, Bacharach [2006, 35-42] calls attention to the practical importance of pure coordination games like Hi Lo (Table 5). Broadly speaking, cooperative and efficient outcomes are more likely when the interests of players tend to converge than when they tend to diverge.[7] A related issue is the degree of "ludic diversity" in evolutionary

6. *Cf.* Hardin [1982, 168-169]: "[The Stag Hunt] is surely not the elemental problem of social order that motivates social contract theory. The elemental problem is how to get someone to be orderly – when it is in his or her interest not to be."

7. Tan and Zizzo [2008] present a measure of *game harmony* – degree of common interest between players – and show experimentally that cooperation is more likely in more harmonious games.

models; as Bacharach [2006, 100] notes, "standard models in bio-evolutionary game theory postulate a ludic ecology having minimal ludic diversity – it contains just one game type." Bacharach [2006, 95-119] sketches some ways in which expanding the set of games played by interactants may impact long-term outcomes; see also Amadae and Lempert [2015].

Second, models can incorporate one-shot games, in which a pair of players interact only once before a new co-player is selected, or repeated games, in which a pair of players engage in repeated interactions before a new co-player is selected. (Usually, the repeated games are of the indefinitely repeated type, where the interaction is terminated after each round (stage game) with some fixed probability.) The repetition of games makes apparent the opportunity for reciprocal interaction, and reciprocity has been long-recognized as a means by which cooperation can emerge and be sustained (*e.g.*, Axelrod [1984]; Trivers [1971]).[8]

A related consideration is the degree of observability in the model: what does a player know about the other player? Most straightforwardly, one cannot implement a reciprocal or conditional strategy without (at least probabilistic) knowledge of other players' moves. Thus, for example, Tit-for-Tat (cooperate in the first round, and then copy co-player's last move) depends on the ability to observe a co-player's immediately prior move. More complex conditional strategies may require a player to observe a longer series of moves or the moves of third parties. *Ex ante* observability of a player's strategy, type, or disposition can also be important. Recent evolutionary models, including Lecouteux ([2015], Chapter 7) and Dekel, Ely and Yilankaya [2007] broadly indicate that – with observability of preferences (types) – efficient and cooperative outcomes can result because it can be evolutionarily advantageous to "advertise" a disposition to cooperate (only) with types who are similarly disposed (see also, in a similar vein, Heller and Winter [2016], discussed below).[9]

The strategies that are included in the evolutionary analysis also can affect outcomes; there are several modeling decisions involved, some with perhaps under-appreciated effects. First, it is relatively clear that *excluding* a strategy by fiat (or inadvertently) may be detrimental to a model's applicability. To take a simple example, if players are randomly drawn to play a repeated PD from a population containing only the types all-C (*i.e.*, always cooperate) and all-D (always defect), in the long-run, only all-D will remain. But, if Tit-for-Tat (play C in first round of interaction, and in subsequent rounds if and only if co-player chose C in the prior round) is also included in the population, the story changes: a long-run outcome in which only all-C and Tit-for-Tat remain is *possible* (*e.g.*, McElreath and Boyd [2007], Chapter 4). So is a long-run outcome in which only all-D remains. In the example, *which* long-run outcome obtains depends on the initial mix of strategies in the population (McElreath and Boyd [2007], 133-135) – this turns out to be an

8. But it is also true that reciprocal strategies do not necessarily lead to efficient outcomes (*e.g.*, Boyd and Richerson [1992]) and that (given certain assumptions) it is possible to sustain cooperation by conditional strategies even if each interaction is with a randomly selected new player, via strategies that punish "innocent" third parties (Kandori [1992]).

9. Binmore [1994, 174-194], in the context of a stylized evolutionary analysis, sounds a cautionary note about the assumption of observable preferences.

important consideration for evolutionary analyses generally, and can especially confound interpretation of simulation-based evolutionary models.[10] Binmore [1998, 315], citing Linster [1990], points out that the relative success of Tit for Tat in Axelrod's [1984] tournament[11] is sensitive to the initial mix of strategies (but see Axelrod [1984], 48). It, and other long-run outcomes, are also potentially sensitive to the existence and characteristics of *mutants* – *i.e.*, strategies not a part of the population at some time $t$ that enter, in small numbers, at $t+1$. Linster [1992] details implications for Alexrod's [1984] tournament. Ultimately, it is not possible to draw *general* conclusions about how these modeling decisions (admissible strategies, initial distribution of types, and mutations) impact the possibility of cooperative and efficient outcomes.

A final consideration is whether players from a population are paired (grouped) to interact randomly, or whether they engage in *assortative pairing* (*assortative interaction*), choosing their co-players. More precisely, although partner choice is often the posited mechanism, the relevant issue is whether pairing is random or not. As I detail further below, non-random pairing has been proposed as a crucial component in the evolutionary viability of cooperation in general (*e.g.*, Sober and Wilson [1998], 23-26, 135-142) and of team reasoning specifically (Bacharach [2006], 101-114); see also closely related discussion in Caporael [2007].

# 3. Interpretation of Evolutionary Models

Before describing a few fundamental results, it is worthwhile to consider the meaning of "evolution" in the context(s) that are relevant to human interaction and team reasoning. (Why) are we entitled to assume that "the strategies of interactants that perform relatively well in terms of some objective payoff at some time $t$ will be relatively better-represented at the next time period $t+1$?" Early evolutionary analyses of cooperation and altruism hypothesized non-human animal actors and, more importantly, modeled *genetic* evolution (see the review in Sober and Wilson [1998], 55-77). The supposition here is that there is a relatively straightforward relationship between an interactant's genotype (specifically, the presence or absence of a particular allele) and its phenotype (an observable expression of its genotype[12] – here, a behavior or "strategy"). If, as assumed, it is true that the phenotype in question affects relative *fitness* – *i.e.*, the payoffs, which are a function of the number of offspring produced – the basic conclusion that

10. For a brief overview of other arguments against simulation in this context, see McElreath and Boyd [2007, 9-11].

11. The tournament was a computer simulation of 63 submitted strategies playing pairwise indefinitely repeated PDs; in a baseline version, each strategy was initially represented in equal proportions, and so each pairing was equally likely.

12. The definition of phenotype is from Wojczynski and Tiwari [2008] which discusses some important subtleties involved in practical applications of the definition.

more successful strategies at time $t$ are better-represented at $t+1$ follows. (In a sense, as Sugden [2001, 121] points out, it is tautological.) If the behaviors that we are concerned with are of the sort assumed in such evolutionary models, straightforward application is unproblematic (for one account of human cooperation defended on these grounds, see Gintis [2000]). Indeed, Aumann [2008, 12], bolsters an evolutionary account by pointing to "a molecular basis for altruism – a real physiological gene" (citing Knafo *et al.* [2008]). However, recent research in genetics, reviewed by Charney and English [2012], casts grave doubt on simple models of association between genotype and even marginally complex behavioral traits (see also Charney and English [2013]; Charney [2012]). More to the point, earlier research indicating substantial (genetic) heritability of such behavior has been called into question (Charney [2012], 331-358, 374-375, 381, 385-392); see also Shultziner [2013*a*], Shultziner [2013*b*].[13]

A second approach is to interpret evolutionary models as models of cultural evolution. Following Richerson and Boyd [2005, 5], define *culture* as "information capable of affecting individuals' behavior that they acquire from [others] through teaching, imitation, and other forms of social transmission," understanding *information* as "any kind of mental state, conscious or not, that can be acquired or modified by social learning and affects behavior," including knowledge, ideas, values, beliefs, and attitudes. A *cultural variant* is one of several pieces of information that may be applied in a given situation.[14]

Cultural evolution is (potentially) a function of several forces, random and non-random (Richerson and Boyd [2005], 68). Particularly relevant are natural selection and biased transmission, of which there are three types. The first, *content-based biased transmission*, entails one cultural variant being learned or retained rather than an alternative, due to a conscious cost-benefit calculation, or relative ease of adoption/recall. The second is *frequency-based biased transmission*, whereby a variant is adopted as a function of its prevalence in the population; the most intuitive example is conformity bias, in which the most common variant in a population is adopted. The final type of biased transmission is *model-based bias*, whereby a cultural variant is adopted because it is exhibited by an individual who is high in some desirable characteristic (*e.g.*, prestige, success, wealth) or is similar in some way to the adopter. *Natural selection* of cultural variants is

13. Charney [2012, 351] describes a striking genome-wide study assessing the genetic correlates of aggression in fruit flies raised in laboratory conditions intended to hold environment constant. At least 266 gene variants were found to be associated with variation in aggression, and heritability of aggression was found to be only about 0.1. One implication is that if the seemingly simple trait of aggression, in a fruit fly, is so far from monogenic, behavior such as human interaction in social dilemmas is bound to be heavily polygenic as well. Similarly, if environment apparently plays such a major role in determining the aggression of laboratory-raised fruit flies, the heritability of complex human behavior is also doubtful. Note that the fruit fly aggression study is not an isolated example; see Charney [2012].

14. In some ways, a cultural variant is analogous to a gene variant (*i.e.*, allele-pair). But competition among cultural variants is different than competition between gene variants. For one, multiple cultural variants can be retained by a single individual (*e.g.*, multiple languages, multiple ways to tie a knot). Still, variants compete for "cognitive resources" of an individual, and also for control of behavior since, typically, in a given situation, one variant will control behavior to the exclusion of others (Richerson and Boyd [2005], 73-76).

closely analogous to genetic natural selection; in brief, cultural variants that "cause people to behave in ways that makes their [variants] more likely to be transmitted increase in frequency (Richerson and Boyd [2005], 76)." (Unlike genetic evolution, of course, transmission frequency here is also affected by the number of people encountered or students tutored, etc.[15])

One means of taking cultural evolution (or simpler modes of learning) into account in evolutionary models has been to import standard models of genetic evolution wholesale, with reinterpretation of their elements (*e.g.* Boyd, Richerson, Gintis and Fehr [2003]; Binmore [1994]; Heller [2015]). Various justifications of this move are given in Bendor and Swistak [1997], Borgers and Sarin [1997], Cabrales [2000] Gale, Binmore and Samuelson [1995], and Schlag [1998]; Grune-Yanoff [2011], Sober [1991], and Sugden [2001] present critiques. Alternatively, scholars have modified standard evolutionary models to incorporate some elements of cultural evolution (or learning) that are analytically distinct from those in genetic evolution; slices of this literature are reviewed in Boyd and Richerson [2010], Friedman ([1998], Appendix B), and Mesoudi, Whiten and Laland ([2006], Section 3.1).

To be a bit more explicit, let $p_j$ be the proportion (relative frequency) of some type $j$ in a population at time period $t$ and let $p'_j$ be that proportion in the time period $t+1$; let $w_j$ be the fitness of type $j$ at $t$ and $\bar{w}$ the average fitness of the population (*i.e.*, the proportion-weighted sum of fitnesses over each type). Then, the standard replicator dynamic (or proportional fitness rule) is given by:

$$p'_j = p_j \frac{w_j}{\bar{w}} \qquad [1]$$

This says that the rate of change in the proportion of a type between two time periods is a linear function of its relative fitness in the first time period (*e.g.* Bendor and Swistak [1998], 108). One might imagine instances of cultural evolution that are consistent with this rule, but some elements of cultural evolution (discussed above) cannot be comfortably fit in this framework. The point here is that the standard replicator dynamic describes a plausible way in which differences in payoffs associated with a type or trait may affect its relative prevalence in the population; however, especially in the case of complex traits, where the genotype-phenotype relationship is not straightforward, the equation may not be the best way to model how payoffs (or, indeed, other attributes of a trait) translate into changes in relative frequency. The standard replicator dynamic applies straightforwardly to models of genetic evolution where there is a tight tie between gene and trait; its simplicity and plausibility have made it a workhorse in evolutionary models in general – still, as I discuss below, there are reasons to be cautious when applying it to model the evolution of team reasoning.

15. Richerson and Boyd [2005, 68] considers three other forces. *Cultural mutation* and *cultural drift* are random processes that have effects roughly akin to those of genetic mutations; *guided variation* is the effect of parents' *learned* behavior that is passed on to offspring (the idea is that learned, small improvements, imitated and then improved-upon by successive generations can lead to cumulative, adaptive change) (Richerson and Boyd [2005], 111-119).

# 4. Basics of Stability and Group Selection

In this section, I briefly describe a few fundamental results that will be relevant to evolutionary analyses of team reasoning. Detailed discussion of these results' relevance to team reasoning is deferred to Section 5.

**Equilibria, Stability, and Evolutionarily Stable Strategies (ESS).** An equilibrium in an evolutionary game exists when the fitness (payoffs) of each type in the population is the same. An equilibrium may be polymorphic (more than one type in the population), as in one-shot Hawk-Dove (Table 3), or monomorphic (only one type in the population), which trivially exists for all games. Often, analysts are concerned with stability of an equilibrium; that is, whether the population returns to the equilibrium state once slightly perturbed.

Smith and Price [1973] and Smith [1974] introduce the notion of an Evolutionary Stable Strategy (ESS). An ESS is a strategy that, if common, cannot be outperformed by a small number of "invading mutants" playing one of a set of alternative strategies. If the ESS outperforms a strategy, it can therefore repel an invasion, given the assumption that strategies that do relatively poorly become less frequent. (An ESS, therefore, entails a stable monomorphic equilibrium.) A bit more specifically, a strategy $i$ is an ESS against an alternative strategy $j$ if (1) $i$ is (weakly) a better reply to itself than is $j$ and (2) if $j$ is also a best-reply to $i$, $i$ must be (strictly) a better reply to $j$ than is $j$.[16] For example, in a one-shot PD (Table 2), D an is ESS against C, but C is not ESS against D. In a one-shot Hi Lo, (Table 5) Hi is ESS against Lo, and Lo is ESS against Hi (given random pairing). In Hawk-Dove (Table 3), Hawk is not ESS against Dove, and Dove is not ESS against Hawk (there is instead a stable internal equilibrium, with the proportion of hawks in the population equaling $v/c$). In a indefinitely repeated PD (with a sufficiently low probability that the game ends in any given round), Tit-for-Tat is ESS against All-D, but not against Tit-for-Two-Tats (Play C in the first two rounds, and then play C if and only if the other played C in one or both of the two previous rounds).

Indeed, no pure strategy is stable against every other pure strategy in an indefinitely repeated PD (Boyd and Lorenbaum [1987]). Still, Axelrod's [1984] claims for Tit-for-Tat as the "best" strategy in the indefinitely repeated PD can, in the main, be supported analytically (Bendor and Swistak [1997]).

**Group Selection.** This section draws heavily on McElreath and Boyd [2007, Ch. 6] and Sober and Wilson [1998, Ch. 1-2]. Consider two groups, where members of each play a series of one-shot PDs, each time with a randomly drawn member of their own group; Group 1 consists predominantly of coop-

16. Symbolically: let $u(x, y)$ be the payoff to $x$ from interacting with $y$; then $i$ is an ESS against $j$ if: (1) $u(i, i) \geq u(j, i)$ and (2) if $u(i, i) = u(j, i)$ then $u(i, j) > u(j, j)$. I have ignored some complications involving mixed strategies, simultaneous invasions, and different means of translating payoffs into subsequent strategy frequencies (*i.e.*, replicator dynamics); see Bendor and Swistak [1998].

erators (always play C), and Group 2 consists predominantly of defectors (always play D). The crude version of the evolutionary argument in favor of group selection says that because the average payoffs of Group 1 members are higher than the average payoffs of Group 2 members, in the long run, Group 1 members (predominantly cooperators) will tend to out-compete Group 2 members (predominantly defectors). What this argument neglects is that *within* each group, defectors are favored, and so will tend to out-compete cooperators. (Even if Group 1 is *all* cooperators, it is vulnerable to an invasion of defectors.) Thus, within-group selection favors individual defectors, but between-group selection appears to favor groups with relatively more cooperators.

The Price equation gives a general formula for the conditions under which a trait will evolve (*i.e.*, increase its proportion in the population) when the population is subdivided into groups (see Price 1972). Note that even a pair of players interacting in a one-shot game can be considered a group in this context. Index players by $i$ and groups by $g$. Consider some trait $T$; let $p$ stand for the proportion of the population with trait $T$ ("$T$-types") in a given time period and $p'$ for that proportion in the subsequent time period, defining $\Delta p \equiv p' - p$. The interesting case is one where $T$-types increase the fitness of fellow group members, relative to members of other groups, but have less fitness than the average group member. Let $n_g$ be the number of $T$-types in group $g$, and $\bar{N}$ the size of the population. Finally, let $w_{ig}\,(>0)$ be the fitness of player $i$ in group $g$, $w_g = \frac{1}{n_g}\sum_i w_{ig}$ be the mean fitness in group $g$, and $\bar{w}$ the mean fitness in the population. Assuming the standard replicator dynamic (proportional fitness rule), it can be shown that:

$$\bar{w}\Delta p = \sum_g \frac{n_g}{N} w_g (p_g - p) + \sum_g \frac{n_g}{N} w_g (p'_g - p_g) \qquad [2]$$

Note that $p$ only increases if this expression is greater than 0, and in the interesting case, the first sum is positive and the second negative. Assuming that noise is negligible, the formula for $\bar{w}\Delta p$ can be further simplified to:[17]

$$\bar{w}\Delta p = \mathrm{cov}(w_g, p_g) + \mathrm{E}[\mathrm{cov}(w_{ig}, p_{ig})] = \mathrm{var}(p_g)\beta(w_g, p_g) + \mathrm{E}[\mathrm{var}(p_{ig})\beta(w_{ig}, p_{ig})], \quad [3]$$

with $\beta(x, y)$ denoting the coefficient from regressing $x$ on $y$. This makes clear that $\Delta p$ increases as (1) $\mathrm{var}(p_g)$, the inter-group variation in proportion of $T$-types, increases; (2) $\beta(w_g, p_g)$, the (positive) association between group fitness and the proportion of $T$-types in the group, becomes greater in mag-

17. The simplification from (2) to (3) involves noting that the right hand side of (2) consists of two expectations, $\mathrm{E}[w_g(p_g - p)]$ and $\mathrm{E}[w_g(p'_g - p_g)]$. The first expectation can be written as $\mathrm{cov}(w_g, p_g)$ since by the definition of covariance, $\mathrm{cov}(w_g, p_g) = \mathrm{E}(w_g p_g) - \mathrm{E}(w_g)\mathrm{E}(p_g) = \mathrm{E}\{w_g[p_g - \mathrm{E}(p_g)]\}$; recall also that by definition $p \equiv \mathrm{E}(p_g)$). The second expectation can be written as $\mathrm{E}(w_g \Delta p_g)$ since by definition $\Delta p_g \equiv p'_g - p_g$. Thus, we have $\bar{w}\Delta p = \mathrm{cov}(w_g, p_g) + \mathrm{E}(w_g \Delta p_g)$. Now, the key move is to realize that this equation implies that $w_g \Delta p_g = \mathrm{cov}(w_{ig}, p_{ig}) + \mathrm{E}(w_{ig}\Delta p_{ig})$ (think for the moment of group $g$ as the population and the individuals within that group as the subpopulation). Finally, invoke the assumption that there is no noise (*i.e.*, that trait T is stable within individuals), which means $\Delta p_{ig} = 0$, to give (3) (see McElreath and Boyd [2007], 228-232).

nitude; (3) $\mathrm{var}(p_{ig})$, the intra-group variation in proportion of $T$-types, decreases; (4) $\beta(w_{ig}, p_{ig})$, the (negative) association between individual fitness and being a $T$-type, decreases in magnitude.

In the case where groups are made up of two players, randomly selected each round from the population to play a one-shot PD,

$$\bar{w}\Delta p = \frac{1}{2}p(1-p)(b-c) - \frac{1}{2}p\ (1-p)(b+c) = p(1-p)(-c)$$

But now suppose that pairing is non-random, and that $T$-types are more likely to pair with other $T$-types – *i.e.*, there is assortative pairing; let the correlation between being a $T$-type and having a $T$-type partner be $r$. Then,

$$\bar{w}\Delta p = \frac{1}{2}p(1-p)(1+r)(b-c) - \frac{1}{2}p\ (1-p)(1-r)(b+c),$$

which, when $T$-types only pair with other $T$-types, is $p(1-p)(b-c)$) (McElreath and Boyd [2007], Ch. 6).

It worth noting, at least in passing, that, for small-scale societies, anthropologists see cultural group selection as much more plausible in practice than genetic group selection. The crux of the reason is that inter-group genetic variance is quite low relative to intra-group variance (ratios $< .05$ are typically cited), since migration among groups serves to make groups genetically similar (McElreath and Boyd [2007], 232-255). But (non-payoff-dependent) migration is less of a factor for the faster-developing cultural evolution, making more plausible group selection for a trait that increases group fitness but is disadvantaged within groups, especially if the within-group selection is relatively weak (for one such model see Boyd *et al.* [2003]). The relevance of this point to team reasoning is that if – as Bacharach [2006, Ch. 3] argues (see below) – group selection is the means by which team reasoning evolves, *cultural* transmission of team reasoning is likely required.

# 5. Evolutionary Analyses of Team Reasoning

Bacharach [2006, Ch. 3] sketches a model that explores the role of team reasoning in human evolutionary history. The model is one of genetic evolution (Bacharach [2006], 96-98), but a potential role for cultural transmission is recognized as well (Bacharach [2006], 119 (fn. 27)). Considering first the one-shot PD, Bacharach [2006, 101-104] argues that group selection can account for the evolution of team reasoning (and thus cooperation); in the analysis, it is recognized that some assortative grouping is required for the evolutionary viability of team reasoning in the PD – *i.e.*, team reasoners (cooperators) must be disproportionately likely to interact with other team reasoners (cooperators). Bacharach [2006, 104-114] makes a second impor-

tant point: the set of games in one's "life game" includes more than the PD. In games like Stag Hunt and Hi Lo, coordination on a Pareto-optimal outcome will be favored by group selection (and, in fact, by individual selection, depending on the game and the makeup of the population). Granting this, there remains the question: is there a mechanism that is evolutionarily favored to bring about such coordinated, cooperative behavior? Bacharach [2006, 111-114] argues that *social identification* with other members of the group – roughly, a tendency to see oneself *as* a group member, and to take the group's goals as one's own – is the mechanism that brings about team reasoning. (On social identity theory generally, see Turner, Hogg, Oakes, Reicher and Wetherell [1987] and Brewer [1991]; in an evolutionary context, see Brewer [2004], Brewer and Caporael [2006], Caporael [1995], Caporael [2007], Caporael and Brewer [1991], Caporael and Brewer [1995], and Caporael and Brewer [2000].) Why social identification? The advantage over other proposed explanations for cooperation (*e.g.*, altruism, reciprocal or otherwise) is that social identification explains a *repertoire* of cooperative actions – a single mechanism can be used in a range of interactions. (Note, for example, that altruism, by itself, does not get interactants to ( Hi ; Hi ) in a Hi Lo.) All else equal, a single mechanism that brings about some set of outcomes will be evolutionarily favored over a set of mechanisms that bring about the same ends (and over a single, more complex mechanism) (see also Sober and Wilson [1998], 304-324). In sum, Bacharach argues that group selection could have brought about a disposition to cooperate in a range of games, from the PD to the Stag Hunt to the Hi Lo, and that group identification, which brings about team reasoning, is the mechanism driving this cooperation.

Amadae and Lempert [2015] considers the long-term viability of team reasoning in a context where group selection cannot operate. The analysis involves a population from which players – who are either individual reasoners or team reasoners – are drawn, randomly, to play one-shot versions of Hi Lo or PD. (The qualitative results hold if Stag Hunt replaces Hi Lo.) In the baseline case, depending on the parameters, team reasoning can be an ESS, individual reasoning can be an ESS, or both can be an ESS. The greater the frequency of games that are Hi Lo relative to the PD, the more successful is team reasoning (*i.e.*, is an ESS and requires a smaller initial proportion of the population to stabilize). In extensions, Amadae and Lempert [2015] suggests that (1) individual reasoners are more prone to certain evolutionarily disadvantageous errors of perception than are team reasoners; (2) team reasoning may be an ESS against a type who switches between modes of reasoning based on the game type, if there are (complexity) costs to switching;[18] (3) both team and individual reasoners may persist in the population indefinitely if either the mix of games played varies over time or individual reasoners "learn" from team reasoners in a certain way; (4) *circumspect* team reasoners (Bacharach [1999]) perform better than unconditional team reasoners against individual reasoners. Amadae and Lempert [2015] also argue that the advantages that a consistent team reasoner has – relative to

18. As Richerson and Boyd [2005, 135] bluntly put a somewhat related point: "all animals are under stringent evolutionary pressure to be as stupid as they can get away with."

individual reasoners and types who switch between modes of reasoning – are particularly likely to be relevant in interactions less straightforward than those encountered in a laboratory setting.

An evolutionary analysis by Lecouteux [2015, Ch. 7] suggests that if players' types are observable by other players, team reasoning will be particularly advantaged, relative to individual reasoning. In the evolutionary model, there are multiple populations; in each time period, one player is randomly drawn from each population to play a one-shot game. Players' types are characterized by *reply functions* that associate a response to each move by a co-player; these reply-functions (or types) are perfectly observable. A "heuristic" is a behavioral rule that selects an action for any distribution of reply-functions. Lecouteux [2015] is interested in determining the characteristics of games for which a population of players with the heuristic of "payoff-maximizing behavior" (*i.e.*, maximizing one's individual, objective payoff – individual reasoning) is evolutionarily stable.[19] Essentially, the conclusion is that payoff-maximizing behavior is certainly *not* an ESS in any game where a set of players can coordinate on an outcome that yields (to every player in the set) a payoff greater than that of the most efficient Nash equilibrium in the game; this coordination is possible because heuristics can incorporate a commitment to coordinate (rather than payoff-maximize) *conditional on* other(s) also coordinating. Additionally, when payoff-maximizing behavior is not an ESS, the type that can invade has a heuristic that is conceptually equivalent to team reasoning.

When considering a player who must make choices in a broad set of games, one may think of team reasoning as a rule prescribing actions that may not be individually optimal in every game, but, as a whole, nonetheless preferable to competing rules of behavior (*e.g.*, individual reasoning). Viewing team reasoning in this perspective, it is worth a slight detour to mention two evolutionary analyses of *rule rationality*.[20] Aumann [2008] proposes the idea of rule rationality and presents an informal evolutionary argument on its behalf. Rule-rationality is contrasted with act-rationality. Act-rationality selects – in a given, narrowly defined scenario – the utility-maximizing *act* from a set of possible acts. (*E.g.*, "should I take my umbrella with me today?" has two possible acts, {take today, don't take today}.) Rule-rationality selects – considering all possible scenarios in a relevant class of decision scenarios – the (approximately) utility-maximizing rule from a set of possible rules; rules assign an act to all scenarios in the relevant class. (For example, a rule might answer the question "When should I carry an umbrella?", and possible rules might include {always, never, when rain chance $> 50\%$,...}.) Utility here is thought of as analogous (though not identical) to fitness in an evolutionary sense (Aumann [2008], 15-17). The interesting situation is when the act implied by rule-rationality and the one implied by act-rationality are not the same; what might justify rule-implied acts in such cases? The informal evolutionary explanation seems to be that the simplicity of broad rules (implying reduced cognitive load) may some-

19. A slightly weaker definition of ESS is used than the "strict" ESS of Smith and Price [1973] discussed above.

20. I thank the editor of this issue for pointing out this literature.

times offset the fact that the acts they imply are not always act-rational (Aumann [2008], 11).

Heller and Winter [2016] considers a "reduced form evolutionary model" consisting of a two-stage game: in the first, players are faced with a set of games from which nature selects one that is played in the second stage. In the first stage, players can commit to be "rule-rational" by bundling a set of games together (*i.e.*, in committing to play the "same move," regardless of which game is realized in stage two). Each player (at least probabilistically) observes the other's commitment (or lack thereof) and by assumption the commitment cannot be broken. In stage two, the players keep their commitment (if made) or maximize their utility in the realized game. Heller and Winter [2016] shows that subgame perfect equilibria exist where commitments to bundle games together are made, thus justifying rule rationality. These results are relevant to team reasoning especially to the extent that "team reasoning" entails a commitment to reason in a specific way across a set of games (for example, because it is induced exogenously by group identification).

# 6. Discussion

Despite the relatively varied approaches that these evolutionary models of team reasoning take, one commonality is the assumption that the standard replicator dynamic governs overtime change in the evolution of types. Following Bacharach [2006, 96-97], it is possible to interpret evolutionary models of team reasoning as models of (genetic) natural selection; this would then justify the use of the standard replicator dynamic. But probably the weight of the evidence suggests, at the minimum, substantial roles for learning and culture (see below for further discussion). Thus, it may be worthwhile to investigate the role that learning rules other than that implied by the standard replicator dynamic might play in the emergence and viability of team reasoning. The literature on cultural evolution (and gene-culture co-evolution) (*e.g.* Richerson and Boyd [2005]) may serve as a fruitful starting point; the dynamics of cultural transmission have been carefully theorized therein, even though modes of reasoning, as a culturally transmitted trait, have not been specifically considered. On the one hand, it is fairly likely that basic results will be robust to varying the assumption of the standard replicator dynamic (Bendor and Swistak [1998]); on the other, some interesting differences may manifest (Cubitt and Sugden [1998]; Sugden [2001]).

These considerations suggest several lines of empirical and theoretical inquiry. Perhaps the overarching question: how *is* team reasoning transmitted? (In the context of cultural evolution: from whom do we learn, and on what basis?) First, one may ask whether team reasoning is relatively stable (*i.e.*, cross-situationally consistent within individuals, and so more conventionally "trait-like"). If, as in Bacharach [2006], group identification is the mechanism that brings about team reasoning, the literature on social identity theory (cited above) provides some hints. Broadly, the literature indi-

cates that social identity is easily primed (even if categories are arbitrarly assigned), and once primed, has reasonably consistent effects across individuals, at least given a specific, well-defined situation (*e.g.*, Kramer and Brewer [1984]; Caporael, Dawes, Orbell and van de Kragt [1989]). At the same time, even given a broadly-shared capacity for social identification, this literature emphasizes that reconciling individual and group goals is a "perpetual juggling act," and that the behavioral implications of social identification differ across group size (Brewer and Caporael [2006]).[21] Experimental evidence in behavioral economics tends to indicate appreciable variability in (behavior consistent with) team reasoning as a function of game type (Colman, Pulford and Lawrence [2014]; Colman, Pulford and Rose [2008]; Tan and Zizzo [2008]), but there is probably still room to explore other elements of trait stability, in particular the extent to which team reasoning is stable within individuals across time.

A question related to stability is whether team reasoning and individual reasoning are appropriately treated as discrete traits. It is at least plausible to treat team reasoning as binary, given a well-defined game at a fixed point in time.[22] But more broadly – especially if team reasoning is understood to be genetically influenced – it is probably more accurate to model the *tendency to team reason* as a continuous trait (Sober and Wilson [1998], 136; in this vein, see also the exogenous $c$ parameter, "representing the willingness of a player to act as part of a [team]" in the model of Butler [2012]). (This is particularly so if multiple games are simultaneously analyzed.) Of course, this may well come at the price of analytical complexity.[23]

There is also a question of how *faithfully* team reasoning, as a putative cultural variant, can be transmitted from person to person. Heinrich, Boyd and Richerson [2008, 121] explains the generic problem:

> [U]nlike genes, ideas are not transmitted intact from one brain to another. Instead, the mental representations in one brain generate observable behavior, a "public representation" [...]. Someone else then observes this public representation, and then (somehow) infers the underlying mental representation necessary to generate a similar public representation. The problem is that there is no guarantee that the mental representation in the second brain is the same as it is in the first. Any particular public representation can potentially generate an infinite number of mental representations in other minds. Mental representations will be replicated from one brain to another only if most people induce a unique mental representation from a given public representation.

21. More precisely, the behavioral implications do not arise only because of group size, but also group function. An evolutionary argument distinguishes four so-called core-configurations: *dyad* (parent-child; intimate partner); *task group* (c. five members, organized to carry out a well-defined task); *deme* (c. 30 members, involved in task group coordination, migration, sharing knowledge); *macrodeme* (c. 300 members, seasonally assembling to exchange persons, resources, and information) (Brewer and Caporael [2006], 150).

22. The theoretical analysis in Smerilli [2012] calls even this narrow claim into question.

23. See Wilson and Dugatkin [1997] for one evolutionary analysis involving a continuous trait of interest. Simulations in Heinrich and Boyd [2002] suggest that the treating continuous traits as discrete (binary) in evolutionary analyses does not greatly influence conclusions drawn.

On the one hand, it is fairly clear that inferring that another is team reasoning *per se* during an interaction is unlikely. True, several experiments (discussed above) have derived means of distinguishing team reasoning from other theories of decision-making that are often observationally equivalent, but the creativity involved in these studies just highlights the difficulty of inferring team reasoning from real-life behavioral interactions. (An obvious but important point is that occasional, or even frequent, observational equivalence between theories does not imply that competing theories have an equal role in driving actual behavior.)

On the other hand, the relatively straightforward nature of team reasoning as an idea (*i.e.*, "ask: what should *we* do") makes it plausible that team reasoning could be transmitted explicitly (taught by a parent to a child for example). As well, Heinrich and Boyd [2002] present three formal models, each representing a plausible variant of cultural evolution, indicating that even transmission that is noisy (*i.e.*, not particularly faithful) need not fundamentally change the conclusions that one would draw from an evolutionary analysis that assumes faithful transmission (but see Claidiere and Sperber [2007]). Still, the extent to which individuals can perceive that someone has engaged in team reasoning per se (and so whether they can learn or adopt team reasoning themselves) is probably a question of inherent interest.[24]

To investigate group selection as a means by which team reasoning spreads (as proposed by Bacharach [2006]), empirical investigations of partner choice may be interesting. To what extent *are* people able to screen others for team reasoning? (How quickly) are individual reasoners excluded from team reasoning groups? There is a connection here also to the question of whether types are observable, which is also theoretically important for evolutionary analyses of team reasoning (Lecouteux [2015], Ch. 7).

It may be possible to bring evidence to bear on something of a divide in the literature, regarding whether team reasoning is a conscious, perhaps strategic, choice. Though evolutionary and "rational-choice" approaches are not inconsistent with each other, as Lecouteux [2015] emphasizes with respect to team reasoning, it is most straightforward to understand evolutionary models as assuming non-strategic actors. At the risk of oversimplification, some scholars tend to emphasize the circumstantial, exogenous antecedents of team reasoning, treating team reasoning (and the perspective or identification that is hypothesized to bring it about) as emerging unconsciously or involuntarily (Amadae and Lempert [2015]; Bacharach [2006]; Colman, Pulford and Rose [2008]; Tan and Zizzo [2008]; Petersson N.d.), while others emphasize that selecting a mode of reasoning can be justified on strategic grounds (Lecouteux [2015]; Smerilli [2012]).[25] Perhaps the post-experiment survey questions in Colman, Pulford and Lawrence

24. And it may be relevant to understanding how individuals' expectations that others *will* team reason (or not) comes about, which is itself relevant to understanding Bacharach's [1999; 2006] $\omega$ parameter. For an overview of complications involved, and a theoretically compelling derivation of $\omega$, see Smerilli [2012].

25. Gauthier [1975] and Gilbert [1989], for example, can also be thought of as proposing that mode of reasoning is a strategic choice; in a similar vein, see also Courtois, Nessah and Tazdait [2015]. For partially overlapping discussion, see Gold and Sugden [2007].

[2014] come closest to assessing explicitly *why*, in a given instance, subjects team reason, but there is clearly room for further work.

Finally, it is worth recalling the cautionary admonition in Bacharach [1993], regarding the players' framing of games. Assuming that the rules of a game are "unmistakable" to players is "extremely restrictive" – particularly for games "in the field" (Bacharach [1993], 257-258). A player's conception of her situation should be carefully theorized – not taken as given (Bacharach [1993], 271). Since any real-life game is subject to players' framing, it may be substantially different from its abstract representation. The further an interaction moves from the confines of the laboratory, the more impact this is likely to have on decision-making.[26]

# 7. Conclusion

A substantial body of literature addresses the evolutionary bases of human interaction and cooperation, and analyses of team reasoning are now among these works. Yet, as I have suggested, there room to consider more carefully the implications that a full-fledged theory of cultural evolution has for these projects. Empirically-oriented scholars might work to ascertain the means by which team reasoning is transmitted, and why it arises in a given interactive situation. Theoretical analyses could probe robustness of results to varying assumptions about what payoffs entail, and how payoffs translate into adoption of team reasoning. Both lines of inquiry have the potential to be as rewarding as they will be challenging.

26. Framing effects may also impact modelers themselves. In a famous case study, Axelrod [1984, Ch. 4] describes World War I trench warfare, arguing that the structure of the interaction was a repeated PD – with "shooting at the enemy" as $D$ and "not shooting" as $C$ – and that a (Tit-for-Tat; Tit-for-Tat) equilibrium, bolstered by soldiers' "relatively clear understanding of the role of reciprocity in maintaining cooperation" explained the each side's "live and let live" approach. However Binmore [1998, 319] – while broadly defending the evolutionary viability of "mean" strategies that start by defecting in repeated games – disputes Axelrod's [1984] assessment that the soldiers played Tit-for-Tat: "[Axelrod's] explanation overlooks the obvious fact that the players did not begin by being nice to each other." Gelman [2008] has a different objection:

> The model's key assumption is that an individual soldier benefits, in the short term, from firing at the enemy. (In the terminology of the prisoner's dilemma, to cooperate is to avoid firing, and the model assumes that, whatever the soldiers on the other side do, you are better off firing, that is, not cooperating.) Thus, elaborate game-theoretic modeling is needed to understand why this optimal short-term behavior is not followed. In fact, however, it seems more reasonable to suppose that, as a soldier in the trenches, you would do better to avoid firing: shooting your weapon exposes yourself as a possible target, and the enemy soldiers might very well shoot back at where your shot came from [internal citation deleted]. If you have no short-term motivation to fire, then cooperation is completely natural and requires no special explanation.

In other words, Gelman [2008] makes a perfectly defensible argument that in the representative stage game (don't shoot; don't shoot) was a (perhaps unique) Nash equilibrium. (Gelman [2008] does not make this point, but one may also question whether in trench warfare, soldiers' discount factors are high enough to sustain cooperation if the stage game has no cooperative equilibrium.) That there is such disagreement among three thoughtful scholars about how to model a relatively well-defined interaction is a reminder that framing effects can impact even the most sophisticated of thinkers.

# 8. Tables

**Table 1. An evolutionary game where, in the long run, only A type that always plays A remains present. Payoffs are listed as (row player; column player).**

| | A | B |
|---|---|---|
| A | 3; 3 | 2; 1 |
| B | 1; 2 | 0; 0 |

**Table 2. Prisoners' Dilemma (PD). $t > r > p > s$. The usual assumption is that $2r > t + s$. Payoffs are listed as (row player; column player).**

| | C | D |
|---|---|---|
| C | $r$ ; $r$ | $s$ ; $t$ |
| D | $t$ ; $s$ | $p$ ; $p$ |

**Table 3. Hawk-Dove. $v, c > 0$. Payoffs are listed as (row player; column player).**

| | Dove | Hawk |
|---|---|---|
| Dove | $\frac{v}{2}$; $\frac{v}{2}$ | 0 ; $v$ |
| Hawk | $v$ ; 0 | $\frac{v-c}{2}$; $\frac{v-c}{2}$ |

**Table 4. Stag Hunt. $s > h > 0$. In another version of the Stag Hunt, the payoff for hunting hare if the co-player hunts stag, denoted $V(H\|S)$, is slightly greater the payoff for hunting hare if the co-player hunts hare, so $V(S\|S) > V(H\|S) > V(H\|H) > V(S\|H)$. Payoffs are listed as (row player; column player).**

| | Stag | Hare |
|---|---|---|
| Stag | $s$; $s$ | 0; $h$ |
| Hare | $h$; 0 | $h$; $h$ |

**Table 5. Hi Lo; $h > l > 0$ Payoffs are listed as (row player; column player).**

| | Hi | Lo |
|---|---|---|
| Hi | $h$; $h$ | 0;0 |
| Lo | 0;0 | $l$; $l$ |

# References


AMADAE S. M. and LEMPERT D. [2015], "The Long-term Viability of Team Reasoning." *Journal of Economic Methodology* 22(4): 462-478.

AUMANN R. J. [2008], "Rule-Rationality Versus Act-Rationality." Center for the Study of Rationality, Hebrew University of Jerusalem, Discussion Paper No. 497.

AXELROD R. [1984], *The Evolution of Cooperation*. New York: Basic Books.

BACHARACH M. [1993], Variable Universe Games. In *Frontiers of Game Theory*, ed. Ken Binmore, Alan Kirman and Piero Tani. Cambridge, MA: The MIT Press p. 255-275.

BACHARACH M. [1999], "Interactive Team Reasoning: A Contribution to the Theory of Cooperation." *Research in Economics* 53(2): 117-147.

BACHARACH M. [2006], *Beyond Individual Choice: Teams and Frames in Game Theory*. Princeton, NJ: Princeton University Press. Edited by Natalie Gold and Robert Sugden.

BARDSLEY N. and ULEA. [2017], "Focal Points Revisited: Team reasoning, the Principle of Insufficient Reason and Cognitive Hierarchy Theory." *Journal of Economic Behavior and Organization* 131(1): 74-86.

BARDSLEY N., MEHTA J., STARMER C. and SUGDEN R. [2010], "Explaining Focal Points: Cognitive Hierarchy Theory Versus Team Reasoning." *The Economic Journal* 120(543): 40-79.

BENDOR J. and SWISTAK P. [1997], "The Evolutionary Stability of Cooperation." *American Political Science Review* 91(2): 290-307.

BENDOR J. and SWISTAK P. [1998], "Evolutioniary Equilibria: Characterization Theorems and their Implications." *Theory and Decision* 45(2): 99-159.

BINMORE K. [1994], *Playing Fair*. Cambridge, MA: MIT Press.

BINMORE K. [1998], *Just Playing*. Cambridge, MA: MIT Press.

BORGERS T. and SARIN R. [1997], "Learning Through Reinforcement and Replicator Dynamics." *Journal of Economic Theory* 77(1): 1-14.

BOYD R. and LORENBAUM J. P. [1987], "No Pure Strategy is Evolutionarily Stable in the Iterated Prisoner's Dilemma Game." *Nature* 327(6117): 58-59.

BOYD R. and RICHERSON P. J. [1992], "Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups." *Ethology and Sociobiology* 13(3): 171-195.

BOYD R. and RICHERSON P. J. [2010], "Transmission Coupling Mechanisms: Cultural Group Selection." *Philosophical Transactions of the Royal Society B* 365(1559): 3787-3795.

BOYD R., RICHERSON P. J., GINTIS H. and FEHR E. [2003], "The Evolution of Altruistic Punishment." *Proceedings of the National Academy of Sciences* 100(6): 3531-3535.

BREWER M. [1991], "The Social Self: On Being the Same and Different at the Same Time." *Personality and Social Psychology Bulletin* 17(5): 475-482.

BREWER M. [2004], "Taking the Origins of Human Nature Seriously: Toward a More Imperialist Social Psychology." *Personality and Social Psychology Review* 8(2): 107-113.

BREWER M. and CAPORAEL L. R. [2006], An Evolutionary Perspective on Social Indentity: Revisiting Groups. In *Evolution and Social Psychology*, ed. Mark Schaller, Jeffrey A. Simpson and Douglas T. Kendrick. New York: Psychology Press p. 143-162.

BUTLER D. J. [2012], "A Choice for me or for us? Using We-reasoning to Predict Cooperation and Coordination in Games." *Theory and Decision* 73(1): 53-76.

CABRALES J. [2000], "Stochastic Replicator Dynamics." *International Economic Review* 41(2): 451-481.

CAPORAEL L. R. [1995], "Sociality: Coordinating Minds, Bodies, and Groups." *Psycoloquoy* 6(1): NP. Accessed at: http://www.cogsci.ecs.soton.ac.uk/cgi/psyc/newpsy?6.01.

CAPORAEL L. R. [2007], Evolutionary Theory for Social and Cultural Psychology. In *Social Psychology: Handbook of Basic Principles*, ed. Arie W. Kruglanski and E. Tory Higgins. 2nd ed. New York: Guildford Press p. 3-18.

CAPORAEL L. R. and BREWER M. [1991], "Reviving Evolutionary Psychology: Biology Meets Society." *Journal of Social Issues* 47(3): 187-194.

CAPORAEL L. R. and BREWER M. [1995], "Hiearchical Evolutionary Theory: There Is an Alternative, and It's Not Creationism." *Psychological Inquiry* 6(1): 30-34.

CAPORAEL L. R. and BREWER M. [2000], "Metaldeology, Evolution, and Psychology: Once More with Feeling." *Psychological Inquiry* 11(1): 23-26.

CAPORAEL L. R., DAWES R. M., ORBELL J. M. and VAN DE KRAGT A. J. C. [1989], "Selfishness Examined: Cooperation in the Absence of Egoistic Incentives." *Behavioral and Brain Sciences* 12(4): 683-739.

CHARNEY E. [2012], "Behavioral Genetics and Postgenomics." *Behavioral and Brain Sciences* 35(5): 331-410.

CHARNEY E. and ENGLISH W. [2012], "Candidate Genes and Political Behavior." *American Political Science Review* 106(1): 1-34.

CHARNEY E. and ENGLISH W. [2013], "Genopolitics and the Science of Genetics." *American Political Science Review* 107(2): 382-395.

CLAIDIERE N. and SPERBER D. [2007], "The Role of Attraction in Cultural Evolution." *Journal of Cognition and Culture* 7(1): 89-111.

COLMAN A. M., PULFORD B. D. and LAWRENCE C. L. [2014], "Explaining Strategic Coordination: Cognitive Hierarchy Theory, Strong Stackelberg, Reasoning, and Team Reasoning." *Decision* 1(1): 35-58.

COLMAN A. M., PULFORD B. M. and ROSE J. [2008], "Collective Rationality in Interactive Decisions: Evidence for Team Reasoning." *Acta Psychologica* 128(2): 387-397.

COURTOIS P., NESSAH R. and TAZDAIT T. [2015], "How to Play Games? Nash Versus Berge Behaviour Rules." *Economics and Philosophy* 31(1): 123-139.

CUBITT R. P. and SUGDEN R. [1998], "The Selection of Preferences Through Imitation." *Review of Economic Studies* 65(4): 761-771.

DEKEL E., ELY J. C. and YILANKAYA O. [2007], "Evolution of Preferences." *Review of Economic Studies* 74(3): 685-704.

FAILLO M., SMERILLI A. and SUGDEN R. [2017], "Bounded Best-response and Collective-optimality Reasoning in Coordination Games." *Journal of Economic Behavior and Organization* 140(1): 317-335.

FRIEDMAN D. [1998], "On Economic Applications of Evolutionary Game Theory." *Journal of Evolutionary Economics* 8(1): 15-43.

GALE J., BINMORE K. J. and SAMUELSON L. [1995], "Learning to be Imperfect: The Ultimatum Game." *Games and Economic Behavior* 8(1): 56-90.

GAUTHIER D. [1975], "Coordination." *Dialogue: Canadian Philosophical Review* 14(2): 195-221.

GELMAN A. [2008], "Methodology as Ideology." *QA Rivista dell'Associazione Rossi-Doria* 19(2): 167-175.

GILBERT M. [1989], *On Social Facts*. London: Routledge.

GINTIS H. [2000], "Strong Reciprocity and Human Sociality." *Journal of Theoretical Biology* 206(2): 169-179.

GOLD N. and SUGDEN R. [2007], Theories of Team Agency. In *Rationality and Commitment*, ed. Fabienne Peter and Hans Bernhard Schmid. Oxford, UK: Oxford University Press p. 280-312.

GRUNE-YANOFF T. [2011], "Evolutionary Game Theory, Interpersonal Comparisons and Natural Selection: A Dilemma." *Biology and Philosophy* 26(5): 637-654.

HAKLI R., MILLER K. and TUOMELA R. [2010], "Two Kinds of We-Reasoning." *Economics and Philosophy* 26(3): 291-320.

HARDIN R. [1982], *Collective Action*. Baltimore, MD: Johns Hopkins.

HEINRICH J. and BOYD R. [2002], "On Modeling Cultural Cognition: Why Replicators are Not Necessary for Cultural Evolution." *Journal of Cognition and Culture* 2(2): 87-112.

HEINRICH J., BOYD R. and RICHERSON P. J. [2008], "Five Misunderstandings About Cultural Evolution." *Human Nature* 18(2): 119-137.

HELLER Y. [2015], "Three Steps Ahead." *Theoretical Economics* 10(1): 203-241.

HELLER Y. and WINTER E. [2016], "Rule Rationality." *International Economic Review* 57(3): 997-1026.

HOLLIS M. [1998], *Trust within Reason*. Cambridge, UK: Cambridge University Press.

HURLEY S. [1989], *Natural Reasons*. New York: Oxford University Press.

KANDORI M. [1992], "Social Norms and Community Enforcement." *The Review of Economic Studies* 59(1): 63-80.

KNAFO A., ISRAEL S., DARVASI A., BACHNER-MELMAN R., UZEFOVSKY F., COHEN L., FELDMAN E., LERER E., LAIBA E., RAZ Y., NEMANOV L., GRITSENKO I., DINA C., AGAM G., DEAN B., BORNSTEIN G. and EBSTEIN R. P. [2008], "Individual Differences in Allocation of Funds in the Dictator Game associated with Length of the Arginine Vasopressin 1a Receptor RS3 Promoter Region and Correlation between RS3 Length and Hippocampal mRNA." *Genes, Brain, Behavior* 7(3): 266-275.

KRAMER R. M. and BREWER M. [1984], "Effects of Group Identity on Resource Use in a Simulated Commons Dilemma." *Journal of Personality and Social Psychology* 46(5): 1044-1057.

LECOUTEUX G. [2015], "Reconciling Normative and Behavioural Economics." PhD Dissertation. École Polytechnique.

LINSTER B. [1990], "Essays on Cooperation and Competition." PhD Dissertation. The University of Michigan.

LINSTER B. [1992], "Evolutionary Stability in the Infinitely Repeated Prisoners' Dilemma Played by Two-State Moore Machines." *Southern Economic Journal* 58(4): 880-903.

MCELREATH R. and BOYD R. [2007], *Modeling the Evolution of Social Behavior: A Guide for the Perplexed*. Chicago: University of Chicago Press.

MESOUDI A., WHITEN A. and LALAND K. N. [2006], "Towards a Unified Science of Cultural Evolution." *Behavioral and Brain Sciences* 29(4): 329-383.

PETERSSON B. N.d., "Team Reasoning and Collective Intentionality." *Review of Philosophy and Psychology*. Forthcoming.

PRICE G. R. [1972], "Extension of Covariance Selection Mathematics." *Annals of Human Genetics* 35(4): 485-490.

PULFORD B. D., COLMAN A. M., LAWRENCE C. L. and KROCKOW E. M. [2017], "Reasons for Cooperating in Repeated Interactions: Social Value Orientations, Fuzzy Traces, Reciprocity and Activity Bias." *Decisio* 4(2): 102-122.

REGAN D. [1980], *Utilitarianism and Co-operation*. Oxford, UK: Oxford University Press.

RICHERSON P. J. and BOYD R. [2005], *Not by Genes Alone*. Chicago, IL: University of Chicago Press.

SCHLAG K. [1998], "Why Imitate, and If So, How?" *Journal of Economic Theory* 78(1): 130-156.

SHULTZINER D. [2013*a*], "Fatal Flaws in the Twin Study Paradigm: A Reply to Hatemi and Verhulst." *Political Analysis* 21(3): 390-392.

SHULTZINER D. [2013*b*], "Genes and Politics: A New Explanation and Evaluation of Twin Study Results and Association Studies in Political Science." *Political Analysis* 21(3): 350-367.

SKYRMS B. [2003], *The Stag Hunt and the Evolution of Social Structure*. Cambridge, UK: Cambridge University Press.

SMERILLI A. [2012], "We-thinking and Vacillation between Frames: Filling a Gap in Bacharach's Theory." *Theory and Decision* 73(4): 539-560.

SMITH J. M. [1974], "The Theory of Games and the Evolution of Animal Conflicts." *Journal of Theoretical Biology* 47(1): 209-221.

SMITH J. M. and PRICE G. [1973], "The Logic of Animal Conflict." *Nature* 246(2): 15-18.

SOBER E. and WILSON D. S. [1998], *Unto Others*. 2nd ed. Cambridge, MA: Harvard University Press.

SOBER E. [1991], Models of Cultural Evolution. In *Conceptual Issues in Evolutionary Biology*, ed. Elliott Sober. 2nd. ed. Cambridge, MA: MIT Press p. 477-492.

SUGDEN R. [1993], "Thinking as a Team: Towards an Explanation of Nonselfish Behavior." *Social Philosophy and Policy* 10(1): 69-89.

SUGDEN R. [2000], "Team Preferences." *Economics and Philosophy* 2(16): 175-204.

SUGDEN R. [2001], "The Evolutionary Turn in Game Theory." *Journal of Economic Methodology* 8(1): 113-130.

TAN J. J. W. and ZIZZO D. J. [2008], "Groups, Cooperation, and Conflict in Games." *Journal of Socio-Economics* 37(1): 1-17.

TRIVERS R. L. [1971], "The Evolution of Reciprocal Altruism." *Quarterly Review of Biology* 46(1): 35-57.

TURNER J. C., HOGG M., OAKES P. J., REICHER S. and WETHERELL M. [1987], *Rediscovering the Social Group: A Self-Categorization Theory*. Oxford, UK: Blackwell.

WILSON D. S. and DUGATKIN L. A. [1997], "Group Selection and Assortative Interactions." *American Naturalist* 149(2): 336-351.

WOJCZYNSKI M. K. and TIWARI H. K. [2008], "Definition of Phenotype." *Advances in Genetics* 60 (Dec 31): 75-105.